DOCUMENT RESUME

ED 421 523 TM 028 848

AUTHOR	Nasser, Fadia; Wisenbaker, Joseph; Benson, Jeri
TITLE	Modeling the Observation-to-Indicator Ratio Using Logistic Regression: An Example from Factor Analysis.
PUB DATE	1998-04-14
NOTE	16p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Correlation; Error of Measurement; *Factor Analysis; Factor Structure; Foreign Countries; Simulation
IDENTIFIERS	*Logistic Regression; Scree Test

ABSTRACT

Logistic regression was used for modeling the observation-to-indicator ratio needed for the standard error scree procedure (SEscree) to correctly identify the number of factors existing in generated sample correlation matrices. The created correlation matrices were manipulated along the number of factors (4,6), sample size (250, 500), magnitude of factor loadings (0.5, 0.8), and degree of interfactor correlations (0, 0.4). Consequently, the observation-to-indicator (n/v) and the indicator-to-factor (v/f) ratios were also changed. The results indicate that the optimal n/v ratio for determining the number of factors by the standard error scree procedure depends on the characteristics of the data. A smaller n/v (7:1) ratio was needed when factor loadings were high and a larger ratio (14-22) was needed with low loading, particularly when factors were correlated. In all conditions, the n/v ratio for the SEscree procedure to correctly identify the true number of factors with high probability exceeded the minimum of 5:1 stated in some of the related literature. Furthermore, the use of logistic regression provided a model for analyzing data from complex simulation studies that makes it very easy to communicate otherwise very complicated relationships. (Contains 4 figures, 3 tables, and 28 references.) (Author/SLD)

Running head: Modeling n/v Ratio

Modeling the Observation-to-Indicator Ratio Using Logistic Regression: An Example from Factor Analysis

Fadia Nasser
Beit Berl College - Israel
Joseph Wisenbaker
The University of Georgia
Jeri Benson
The University of Georgia

Paper presented at the annual meeting of the American Educational Research Association, San Diego, California, April 14, 1998

Abstract

Logistic regression was used for modeling the observation-to-indicator ratio needed for the standard error scree procedure (SEscree) to correctly identify the number of factors existed in generated sample correlation matrices. The created correlation matrices were manipulated along the number of factors (4, 6), sample size (250, 500), magnitude of factor loadings (.5, .8), and degree of interfactor correlations (0, .4). Consequently the observation-to indicator (n/v) and the indicator-to-factor (v/f) ratios were also changed. The results indicated the optimal n/v ratio for determining the number of factors by the standard error scree procedure depends on the characteristics of the data. A smaller n/v (7:1) ratio was needed when factor loadings were high and larger ratio (14-22) was needed with low loading particularly when factors were correlated. In all conditions, the n/v ratio for the SEscree procedure to correctly identify the true number of factors with high probability exceeded the minimum of 5:1 stated in some of the related literature. Furthermore, the use of logistic regression provided a model for analyzing data from complex simulation studies that makes it very easy to communicate otherwise very complicated relationships.

Key Words: Logistic Regression, Standard Error Scree, Factor Analysis, Number of factors Hit Rate, Observation-to-Indicator Ratio

The number of observations needed to arrive at dependable conclusions has always been an issue in scientific research. The performance of most statistical methods is directly or indirectly affected by sample size because precision of estimate is largely affected by sample size (Pedhazor & Pedazhor, 1991). Approaches to the determination of sample size include general and specific formulas to given sampling designs or research (e.g., Cochran, 1977; Hays, 1988; Jaeger, 1984; Kirk' 1982; Kish, 1965; Winer, 1971), tables (e.g., Rotton & Schönemann, 1978; Tiku, 1967), and power functions charts (e.g., Pearson & Hartley, 1951). A more extensive treatment is found in Cohen (1988) and in Kraemer and Theimann (1987). Cohen (1988) provided a detailed presentation of the elements of power analysis and illustrative application in diverse research contexts such as t-test, correlation, analysis of variance, and multiple regression. In the aforementioned treatments of sample size the number of indicators/predictors was not rigorously dealt with

The determination of the sample size in relation to the number of indicators was also treated by several researchers and they have offered rules of thumb by which to determine the sample size. Ad hoc rules of thumb for statistical models such as multiple regression suggest the number of observations to the number of indicators ratio should be 10:1 to deal with problems of sampling variability and to ensure reasonable power (Tanaka, 1987). Huberty (1994), in the context of discriminant analysis, suggested that in order to estimate hit rates validly, the minimum number of observations in the smallest group should be at least three to five times the number of predictors, conditional on the type of discriminant analysis employed.

In the context of factor analysis, despite the general agreement that large samples are imperative for stability of factor analytic results, there is no agreement as to what constitutes large. For example Cattell (1978) referred to samples below 200 as "smallish" (p. 492). Comrey (1978) recommended a sample of at least 200 observations, but he added that 2000 observation were needed to stabilize the factor structure. Several rules of thumb also have been suggested. Among these is Cattell's (1952) 4:1 observations-to-indicator (n/v) rule. Nunnally (1978) suggested that "a good rule is to have at least 10 times as many subjects as variables" p. 421". Gorsuch (1983) suggested five to ten observations per indicator, or several hundreds. Cliff (1987) offered rather looser guidelines. He stated "with 40 or so variables, a group of 150 persons is about minimum, although 500 is preferable" p. 339).

Wolins (1982) rightly inscribed such rules of thumb as "incorrect" (p.64) solutions to the sample size, as it depends on specific objectives of the analysis and the characteristics of the data. According to Wolins (1982), the required sample size varies, depending, among other things, on the number of factors expected; whether or not it is "necessary to obtain good estimates of individual factor " (p.64); whether or not the indicators are "well behaved" (p.64); and magnitudes of correlation among the indicators.

There is no doubt that the situation is complex and cannot be resolved by simple answers and the suggestion to have as large a sample as feasible that frequently cited by researchers is not always useful. Thus, how large should the sample size be and what is the trade off among sample size, number of indicators, degree of loading and other characteristics of the data?. This study was performed to address this question..

Purpose of the study

The purpose of the study was to determine the optimal (n/v) ratio needed to correctly identify the number of common factors by the standard error scree procedure. The standard error scree procedure was the optimal method for determining the number of factors among the regression-based variations of the visual scree examined in previous work (Nasser, 1997).
In addition the study aimed to provide an example of using a modeling approach such as logistic regression to facilitate the interpretation of extensive and complex research results that are not easy to explicate otherwise.

Method

Design and data generation

One hundred sample correlation matrices were generated from population correlation matrices using Kaiser and Dickman's (1962) method. The standard error scree procedure (Zoski & Jurs, 1996) was used to determine the number of factors incorporated in the created correlation matrices. The created correlation matrices were manipulated along the number of factors (4, 6), sample size (250, 500), magnitude of factor loadings (.5, .8), and degree of interfactor correlations (0, .4). The design is not completely crossed (see Appendix A). Although the n/v was not an explicit variable in the design, it was changed by manipulating sample size and number of indicators. The levels of the manipulated variables were chosen to be sufficiently different in order that their effect on the performance of the standard error scree procedure would be clear, and to keep the design to a manageable level. The correlation matrices were created based on the common factor model as proposed by Gorsuch (1983). The data generation was performed via IML/SAS procedure for PC, Release 6.08.

Data analysis

The number of factors determined by the standard error scree procedure for each of the 100 samples under each condition were aggregated in one data file. The percentage of time the SEscree procedure indicated the true number of factors (hit rate) was computed. Then logistic regression was performed to obtain the predicted probabilities of determining the true number of factors. Predictive discriminant analysis is a useful alternative method to address the same question, However, some of the literature in which the two methods were contrasted indicate that although the two methods yielded similar results, logistic regression is based on fewer assumptions, is more robust with respect to violations of assumptions, is easier to interpret, and is more parsimonious where relevant (e.g., Aldrich & Nelson, 1984; Cleary & Angel, 1984; Dattalo, 1994; Shott, 1991).

The dependent variable (Y) in the logistic regression was a dichotomous variable with a value of one when the standard error scree indicated the true number of factors and zero when the procedure indicated an incorrect number of factors. The predictor variable in the regression model was n/v ratio. Four logistic equations were obtained under four different combinations of the degree of factor loadings and interfactor correlation. In each of the four regression equations the n/v ratio was the predictor variable.

A set of predicted probabilities were obtained by the regression equations under each of the four conditions. Each set of predicted probabilities was plotted against the n/v values included in the study and the optimal n/v ratio for determining the true number of factors under each condition was calculated. The actual probabilities of obtaining the correct number of factors by the standard error scree were plotted on the same ordinates to demonstrate the degree of match between the predicted probability by the model and the actual probability obtained by the standard error scree procedure.

Results

Inspection of the hit rates in Table 1 indicated the sensitivity of the SEscree procedure to the n/v ratio was conditional on the degree of loadings, and to some extent on the degree of

Insert Table 1 Here

correlation among the factors. The hit rates are the actual probabilities of the SEscree to correctly identify the true number of factors. To facilitate the interpretation of the results in Table 1, four logistic equations were obtained under the four different combinations of degree of loadings and interfactor correlation and a summary of the logistic regression estimates are provided in Table 2.The statistically significant b's along with the large R^2_L, in particular when

Insert Table 2 Here

the loadings were .8, indicated that the n/v ratio is strongly related to the probability of the procedure indicating the correct number of factors. The strength of this relationship decreased when the loadings decreased, and especially so when the factors were correlated.

The set of predicted probabilities obtained by the regression equations under each of the four conditions summarized in Table 2, was plotted against the n/v ratio values included in the study. The actual probabilities of obtaining the correct number of factors by the SEscree (see Table 1) were plotted on the same ordinates to demonstrate the degree of match between the predicted probability by the model and the actual probability obtained by the procedure.

Figures 1-4 describe the relationship between the predicted and the actual probability to determine the correct number of factors by the SEscree procedure and the n/v ratio. Figure 1 and 2 indicated that when the loadings were .8 and the factors were correlated or uncorrelated, seven

Insert Figure 1 and 2 here

observations per indicator were needed for the standard error procedure to have perfect probability of indicating the correct number of factors. Nonetheless, when the factors were uncorrelated predicted and actual probabilities completely matched, while a slight discrepancy between the actual and predicted probabilities was noticed when the factors were correlated.

As indicated by Figure 3, with factor loadings of .5 and uncorrelated factors, at least 14 observations per indicator were required to insure that the predicted probability was higher than .8, and 16 observations per indicator were needed to obtain perfect predicted probability.

Insert Figure 3 here

When the factors were correlated at .4 and loadings were .5 (Figure 8), larger n/v ratios were required to obtain high predicted probability. At least 18 observations per indicator were needed to obtain a probability of .8 or higher, and 22 observations per indicator were needed to obtain perfect predicted probability. In the last two Figures the discrepancy between the actual

Insert Figure 4 here

and predicted probability was apparent, in particular in Figure 4.

Discussion and conclusions

Modeling the number the of observation per indicator by logistic regression results indicated that the optimal n/v ratio for determining the number of factors by the standard error scree procedure depends on the characteristics of the data. A smaller n/v ratio was needed when the factors were uncorrelated and factor loadings were high. In all conditions, the n/v ratio for the SEscree procedure to correctly identify the true number of factors with high probability exceeded the minimum of 5:1 stated in some of the related literature (e.g., Cattell, 1952; Cliff, 1987; Gorsuch, 1983). The results of the current study supported Wolins' (1982) argument that the existing rules of thumb concerning the observation-to-indicator ratio are "incorrect" because they ignore the data characteristics.

The major conclusion from modeling the n/v ratio is that general guidelines are usually useless and often misleading. Perhaps providing a general set of guidelines for all data situations is not a wise practice. The results of this study showed specific characteristics of the data tended to require more observations to variables than generally recommended. Therefore guidelines need to be given in the context of other variables and should be situation specific.

This study provides concrete and useful guidelines regarding the optimal n/v ratio needed when applying the standard error scree for determining the number of factors to retain.

There is a reason to believe that the conclusions from this study may hold true with different data analysis approaches, such as multiple regression, discriminant analysis and others. Therefore, this study should be viewed as a pioneer and stimulus for investigating the issue of n/v more intensively and with other statistical approaches. Furthermore, the use of logistic regression to capture information about relationships among the independent variables of interest and whether the correct number of factors are identified provides a model for analyzing data from complex simulation studies that makes it very easy to communicate otherwise very complicated relationships.

References

Aldrich, J. H., & Nelson, F. D. (1984). Linear probability logit, and probit models. Newbery Park, CA, Sage.

Cattell, R. B. (1952). Factor analysis: An introduction and manual for the psychologist and the social scientist. New York: Harper & Brothers.

Cattell, R. B. (1978). The scientific use of factor analysis in behavioral sciences. New York: Plenum.

Cleary, P. D., & Angel, R. (1984). Analysis involving dichotomous dependent variables. Journal of Health and Social Behavior, 25, 334-348.

Cliff, N. (1987). Analyzing multivariate data. San Diego: Hartcourt Brace Jovanovich.

Cochran, W. G. (1977). Sampling Techniques (3rd ed.). New York: Wiley.

Cohen, J. (1988). Statistical power analysis for behavioral sciences (2nd ed.). Hillsdale, Nj: Erlbaum.

Comrey, A. L. (1978). Common methodological problems in factor analytic studies. Journal of Counseling and Clinical Psychology, 46, 648-659.

Dattalo, P. (1994). A comparison of discriminant analysis and logistic regression. Journal of Social service Research, 19, 121-144.

Gorsuch, R. L. (1983). Factor Analysis (2nd ed.). Hillsdale, NJ: Erlbaum.

Hays, W. L. (1988). Statistics (4th ed.). New York: Holt, Rinehart & Winston.

Huberty, C. J. (1994). Applied discriminant analysis. New York: Wiley.

Jaeger, R. M. (1984). Sampling in Education and the social science. New York: Longman.

Kaiser, H. F., & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from arbitrary population correlation matrices. Psychometrika, 27, 179-182.

Kirk, R. E. (1982). Experimental design: Procedures for the behavioral sciences (2nd ed.). Belmont, Ca: brook/Cole.

Kish, L. (1975). Representation, randomization, and control. In H. M. Blalock, A. Aganbegian, F. M. Borodkin, R. Roudon, & V. Capecchi (Ed.), Quantitative sociology. International perspectives on mathematical and statistical modeling (pp. 261-284). New York: Academic Press.

Kraemer, H. C., & Thiemann, S. (1987). How many subjects? Statistical Power analysis in research. Newbury Park, CA: Sage.

Nasser, F. (1997). The performance of regression-based variations of the scree procedure for determining the number of common factors. Unpublished doctoral dissertation, University of Georgia, Athens, GA.

Nunnally, J. C. (1978). Psychometric theory (2nd ed.). New York: McGraw-Hill.

Pearson, E. S., & Hartley, H. O. (1951). Charts of power function for analysis of variance tests, derived from non-central F distribution. Biometrika, 38, 112-130.

Pedhazur, J. E., & Pedahzur-Schmelkin, L. (1991). Measurement, design, and analysis : An integrated approach. Hillsdale, NJ: Erlbaum.

Rotton, J., Schönemann, P. H. (1978). Power tables for analysis of variance. Educational and Psychological Measurement, 38, 213-229.

Shott, S. (1991). Logistic regression and discriminant analysis. Journal of the American Veterinary Medical Association (JAVMA), 198, 1902-1905.

Tanaka, S. T. (1987). "How big is big enough?": Sample size to the number and goodness of fit in structural equation models with latent variables. Child Development, 58, 134-146.

Tiku, M. L. (1967). Tables of power of the F test. Journal of American Statistical Association, 62, 525-539.

Winer, B. J. (1971). Statistical principles in experimental design (2nd ed.). New York: McCraw-Hill.

Wolins, L. !982). Research mistakes in social and behavioral sciences. Ames, IA: Collin, Iwa State University Press.

Zoski, K., & Jurs, S. (1996). An objective counterpart to visual test for factor analysis: The standard error scree. Educational and Psychological Measurement, 56, 443-451 .

Appendix A

Table A-1

Design Variables for the Sample Data

Orthogonal Factors					Oblique Factors (r=.4)				
f	n	v	v/f	l	f	n	v	v/f	l
4	250	16	4:1	.5	4	250	16	4:1	.5
				.8					.8
		36	9:1	.5			36	9:1	.5
				.8					.8
		48	12:1	.5			48	12:1	.5
				.8					.8
	500	16	4:1	.5		500	16	4:1	.5
				.8					.8
		36	9:1	.5			36	9:1	.5
				.8					.8
		48	12:1	.5			48	12:1	.5
				.8					.8
6	250	36	6:1	.5	6	250	36	6:1	.5
				.8					.8
		48	8:1	.5			48	8:1	.5
				.8					.8
	500	36	6:1	.5		500	36	6:1	.5
				.8					.8
		48	8:1	.5			48	8:1	.5
				.8					.8

Note. f=number of factors, v=number of indicators, and l=loading size

Table 1

Percentage of Times(Hit Rate) the SEscree Procedure Determined the Correct Number of Factors under each Condition

Variable						SEscree	
f	n	v	v/f	n/v	l	r=0	r=.4
4	250	16	4:1	15.6	.5	98	59
					.8	100	100
		36	9:1	6.9	.5	4	1
					.8	99	98
		48	12:1	5.2	.5	0	0
					.8	7	4
	500	16	4:1	15.6	.5	100	100
					.8	100	100
		36	9:1	6.9	.5	83	77
					.8	100	100
		48	12:1	5.2	.5	2	4
					.8	100	99
6	250	36	6:1	6.9	.5	9	7
					.8	98	99
		48	8:1	5.2	.5	0	0
					.8	9	8
	500	36	6:1	6.9	.5	82	79
					.8	100	100
		48	8:1	5.2	.5	4	6
					.8	100	100

Note. r=correlation among the factors f=number of factors. v=number of indicators v/f=indicator-to-factor ratio, l=degree of loading. SEscree=standard error scree procedure.

Table 2

Results of the Logistic Regression with Observation-to-Indicator ratio as a Predictor Variable for The Standard Error Scree Procedure

Condition	Intercept	b	P	R^2_L
l=.8, r=0	-22.32	3.82	0.000	.85
l=.8, r=.4	-20.76	3.49	0.000	.84
l=.5, r=0	-11.40	0.92	0.000	.68
l=.5, r=.4	-5.70	-5.70	0.000	.59

Note. l= degree of factor loadings, r=interfactor correlation.

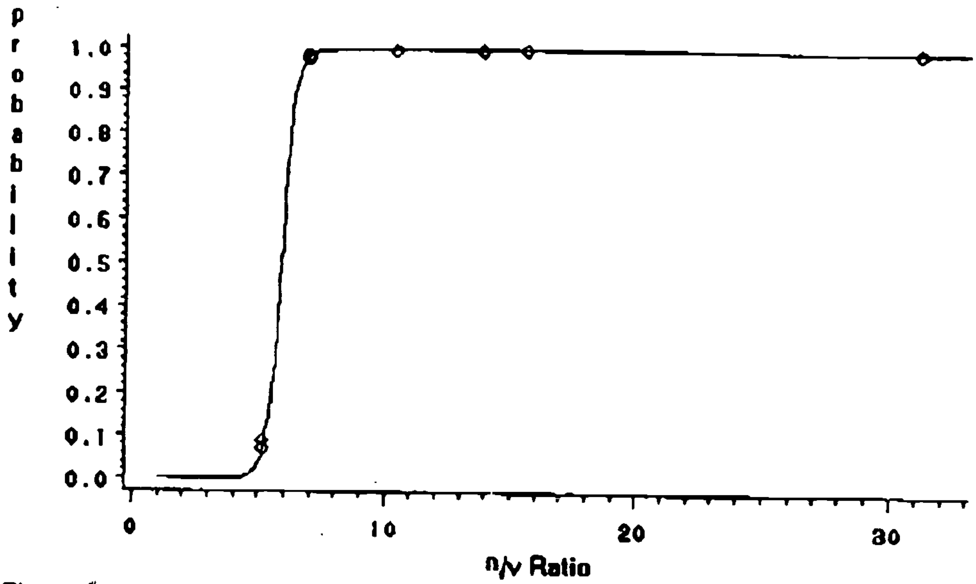

Figure 1. Probability of determining the correct number of factors by the standard error scree procedure as a function of the n/v Ratio (l=.8, r=0).

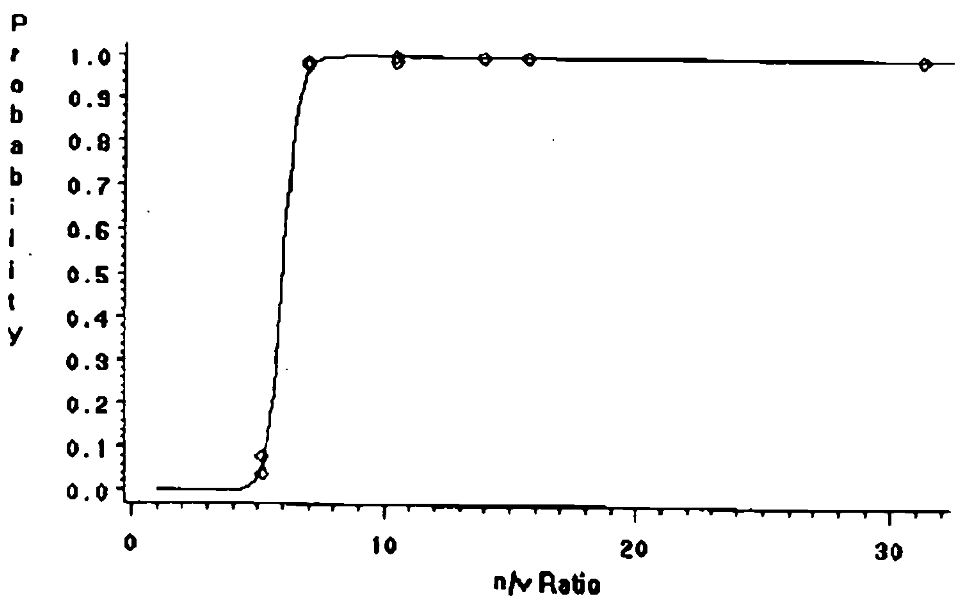

Figure 2. Probability of determing the correct number of factors by the standard error scree procedure as a function of the n/v ratio (l=.8, r=.4).

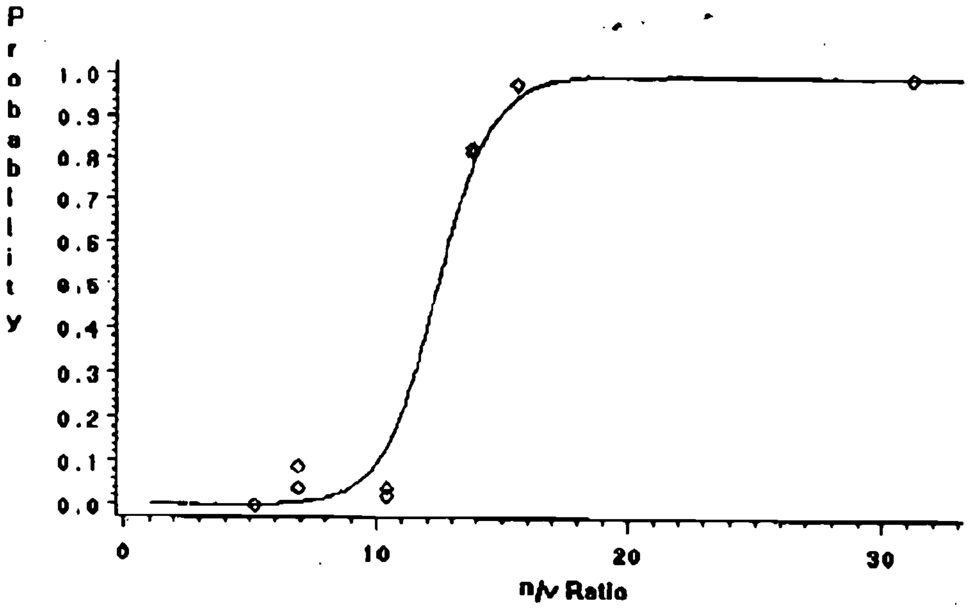

Figure 3. Probability of determining the correct number of factors by the standard error scree procedure as a function of the n/v ratio (l=.5, r=0).

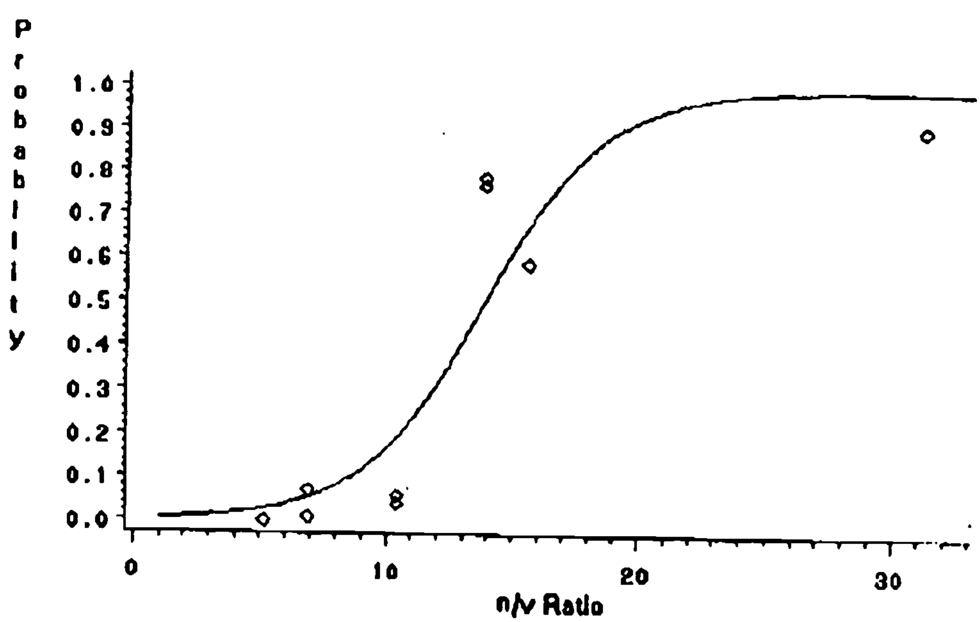

Figure 4. Probability of determining the correct number of factors by the standard error scree procedure as a function of the n/v ratio (l=.5, r=.4).

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM028848

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Modeling the Observation-to-Indicator Ratio Using Logistic Regression: An Example from Factor Analysis

Author(s): Fadia Nasser; Joseph Wisenbaker & Jeri Benson

Corporate Source:

Publication Date: AERA

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1	Level 2A	Level 2B
☑	☐	☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Fadia Nasser	Printed Name/Position/Title:	
Organization/Address:	Telephone:	FAX:
	E-Mail Address:	Date:

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

THE UNIVERSITY OF MARYLAND
ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
1129 SHRIVER LAB, CAMPUS DRIVE
COLLEGE PARK, MD 20742-5701
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com